Neel Sarkar

17 Ross St., Toronto, ON M5T 1Z8
sohail.sarkar@utoronto.ca (437) 329 – 3448 n33levo@github sohail-sarkar@linkedin

Education

University of Toronto **exp. April 2027**
B.Sc. Applied Mathematics (Honours); Minor CS *Toronto, ON*
Trinity College **June 2021**
BA Music, ACTL Classical Guitar *Toronto, ON*

Work Experience

Applied AI/Econometrics Intern June 2025 – Present
Sigma Squared, EO Ventures, Boston, MA

- Engineered agentic EDA–ETL (CatBoost+TabPFN) predicting SWE-error, var -27%, & RMSE -8% (5-fold CV).
- Developed Agentic SDK with MCP wrapping, scaling 50+ workflows across research pipelines.
- Co-authoring proprietary research on econometric-based performance clustering and calibration.

Data Science Intern May – July 2025
PSP Investments, Montréal, QC

- Architected alpha-signal ETL pipeline answering 3,000+ analyst questions per batch.
- Reinforced post-processing with claimification-based cross-referencing; achieved 100% factual accuracy.
- Built Scala retriever layers for preprocessing & learned query routing; cut batch inference 20→3 min.

MLOps Intern January – April 2025
University of Toronto, Toronto, ON

- Deployed RAG chatbot; CI/CD automaton for indexing, evals & releases, cut release lead-time by 60% (5d→2d).
- Delivered a distributed, hierarchical retriever supporting multi-tenant chat.
- Exposed library GPU cluster with REST control-plane for LLM fine-tuning with job scheduling and telemetry.

Software Engineering Intern Sept – Dec 2024
E.J. Pratt Institute, Toronto, ON

- Developed version control scheduling; optimized Random Forest allocator; 6× faster allocation & 12% more coverage.
- Shipped OAuth-secured ticketing (MongoDB + Redis); reduced IT desk resolution time by 40% (3d→1.8d).

Projects

Personal Website: https://n33levo.github.io; AI Representative: https://n33levo.github.io/ai

Fine-Tuned Stock Analysis LLM Nov 2024
QLoRA, PyTorch, BitsAndBytes, PEFT, CUDA

- Fine-tuned Llama-3.1-8B with 4-bit QLoRA (bitsandbytes/PEFT) on market reports, earnings-call transcripts, & SEC filings (10-K/10-Q); applied Bayesian rank-gating to adapt layer ranks, kept quality while reducing compute & memory.

Recommendation Engine Mar 2024
Neo4j, PyTorch, FastAPI, Dagster, AWS EKS(Kubernetes), Docker

- Hybrid recommender: implicit-feedback ALS on watch history & Neo4j knowledge-graph features with a trained re-ranker; Dockerized FastAPI inference & Dagster refresh pipelines and deployed on EKS with autoscaled endpoints.

OCR Digitizer Desktop Software May 2023
TensorFlow.NET, C#, , Keras.NET, PostgreSQL, Pandas/Excel

- OCR that transforms scans into structured data via CRNN(CTC, beam search) w/ layout segmentation & data cleanup(de-dup, schem-val, norm); stores in PostgreSQL, exports to Excel; on-prem license, CAD $35K revenue.

Modified SIR Feb 2022
immunology, PDEs, SciPy, Pandas, PyMC3, FilterPy

- Age-stratified SEIRV extension of SIR (Exposed/Vaccinated, heterogeneous risk); solved ODEs in SciPy with Bayesian fit in PyMC3; applied Unscented & Particle Filters (FilterPy) for sequential updates of S/E/I/R/V on COVID-19 data.

Skills

Python, Go, C++, SQL, PyTorch, Hugging Face, LlamaIndex, LangChain/LangGraph, (Q)LoRA, RAG, vLLM, Triton, Ds.py, PCA, FastAPI, FastMCP, Docker, Kubernetes, AWS, Spark, Databricks, MLflow, Jenkins, REST, GraphQL, Neo4j, SQLite, MongoDB, Streamlit